DOCUMENT RESUME

ED 463 302 TM 033 729

AUTHOR Flowers, Claudia P.; Raju, Nambury S.; Oshima, T. C.
TITLE A Comparison of Measurement Equivalence Methods Based on Confirmatory Factor Analysis and Item Response Theory.
PUB DATE 2002-04-00
NOTE 28p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Item Response Theory; *Measurement
IDENTIFIERS *Confirmatory Factor Analysis; Linear Measurement; Type I Errors; Type II Errors

ABSTRACT

Current interest in the assessment of measurement equivalence emphasizes two methods of analysis, linear, and nonlinear procedures. This study simulated data using the graded response model to examine the performance of linear (confirmatory factor analysis or CFA) and nonlinear (item-response-theory-based differential item function or IRT-Based DIF) methods. Two CFA procedures, lambda (slope) structure only and lambda and tau (slope and intercept) structures, were used to examine measurement equivalence across focal and reference groups. An IRT-based, noncompensatory DIF (NC-DIF) procedure (N. Raju, W. van der Linden, and P. Fleer, 1995) was also used to examine measurement equivalence across groups. Results indicate that the lambda procedure successfully identified items that had differences in the alpha-parameters, but did not identify items that had differences in the beta-parameters. The lambda/tau and NC-DIF procedures identified items that had differences in the beta-parameters. The lambda-tau and NC-DIF procedures were not, however, as sensitive to items that had differences only in the alpha-parameters. When the focal and reference groups had different ability distributions (or impact), the lambda/tau procedure had a lower (or an acceptable) Type II error rate (in detecting true positives), but had a much higher (or an unacceptable) Type I error rate (in detecting false positives). The NC-DIF procedure appeared to have acceptable Type I and Type II error rates in both no-impact and impact scenarios. (Contains 5 tables and 23 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 463 302

RUNNING HEAD: Measurement Equivalence Methods

A Comparison of Measurement Equivalence Methods Based on Confirmatory Factor Analysis and Item Response Theory

Claudia P. Flowers

The University of North Carolina at Charlotte

Nambury S. Raju

Illinois Institute of Technology

T. C. Oshima

Georgia State University

TM033729

Paper presented at NCME Annual Meeting, New Orleans, May 2002

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

C. Flowers

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Abstract

Current interest in the assessment of measurement equivalence emphasizes two methods of analysis, linear and non-linear procedures. This study simulated data using the graded response model to examine the performance of linear (CFA) and non-linear (IRT-based DIF) methods. Two CFA procedures, lambda (slope) structure only and lambda and tau (slope and intercept) structures, were used to examine measurement equivalence across focal and reference groups. An IRT-based, NC-DIF procedure (Raju, van der Linden, & Fleer, 1995) was also used to examine measurement equivalence across groups. Results indicated that the lambda procedure successfully identified items that had differences in the *a*-parameters, but did not identify items that had differences in the *b*-parameters. The lambda/tau and NC-DIF procedures identified items that had differences in the *b*-parameters. The lambda/tau and NC-DIF procedures were not, however, as sensitive to items that had differences only in the *a*-parameters. When the focal and reference groups had different ability distributions (or impact), the lambda/tau procedure had a lower (or an acceptable) Type II error rate (in detecting true positives), but had a much higher (or an unacceptable) Type I error rate (in detecting false positives). The NC-DIF procedure appeared to have acceptable Type I and Type II error rates in both no-impact and impact scenarios.

A Comparison of Measurement Equivalence Methods Based on Confirmatory Factor Analysis and Item Response Theory

A test or subscale is said to have measurement equivalence across groups or populations if persons with identical scores on the latent construct have the same expected raw score or true score at either the item level, the subscale total score level, or both (Drasgow & Kanfer, 1985). Currently, the assessment of measurement equivalence emphasizes two methods of analysis – a linear method (based on confirmatory factor analysis) and a nonlinear method (differential item and test functioning using item response theory). Both approaches test for the equality of true scores across two populations when the latent score is held constant. Recently, a few studies (Raju, Laffitte, & Byrne, 2000; Reise, Widaman, & Pugh, 1993) have offered a comparison of results from these two approaches. The purpose of this study was to examine these measurement equivalence methods under stimulated conditions.

Measurement Equivalence Methods

The following section provides an overview of the confirmatory factor analysis (CFA) and differential item functioning (DIF) using item response theory methods for examining measurement equivalence. Only the single underlying factor CFA model and unideminsional IRT model will be presented. For a more detailed explanation of CFA and DIF measurement equivalence methods, refer to Raju, Laffitte, and Byrne (in press).

Confirmatory Factor Analysis

The linear confirmatory factor analysis (CFA) model may be represented as:

$$x = \Lambda_x \xi + \delta, \qquad (1)$$

where x represents a vector of q x 1 observed variables, ξ is a vector of n x 1 latent variables, Λ_x is a q x n regression coefficients or factor loading matrix that relates n factors to q observed

variables, and δ is a q x 1 vector of measurement errors/residuals in x. Equation 1 is commonly referred to as the measurement model for the exogenous variables in structural equation modeling (SEM) (Jöreskog & Sörbom, 1996).

Within the CFA one-factor model, the underlying construct and the item score can be expressed as (Bollen, 1989; Byrne, 1994):

$$x_i = \lambda_i \xi + \delta_i , \tag{2}$$

where x_i is the observed score on item i, λ_i is the factor loading for item i, and δ_i is the residual/error for item i. The expectation (ε) and the variance of x_i can be expressed as:

$$\varepsilon(x_i) = \lambda_i \xi , \tag{3}$$

$$\sigma_{x_i}^2 = \lambda_i^2 \sigma_\xi^2 + \sigma_{\delta_i}^2 . \tag{4}$$

It should be noted that in Equations 2 and 3, as well as in Equation 1, the intercept is assumed to be zero; that is, the mean of x_i and the mean of ξ are assumed to be zero. This assumption is typically made in CFA, and it will be relaxed subsequently in order to assess measurement equivalence within the context of mean and covariance structure (MACS) analysis of Sörbom (1974).

Using Equation 1, the expectation and the variance-covariance matrix among the observed variables (Σ_x) can be expressed as:

$$\varepsilon(x) = \Lambda_X \xi , \tag{5}$$

$$\Sigma_x = \Lambda_x \Phi \Lambda_x' + \Theta_\delta , \tag{6}$$

where Λ_x' is the transpose of Λ_x, Φ is a variance-covariance matrix of factor variables (ξ), and Θ_δ is a diagonal matrix of measurement error variances. CFA usually places several

restrictions on the elements of the factor-loading matrix (Λ_x), whereas exploratory factor analysis (EFA) does not place any restrictions on the elements of the factor-loading matrix (Λ_x).

Testing of equivalence across populations. In assessing the existence of a common factor model across populations, several invariance tests are typically performed. A test of equality of Σ (observed variance-covariance matrices) across populations has been recommended before testing for invariance of other parameters (Jöreskog & Sörbom, 1989). However, Byrne (1998) reported that it is possible to have invariance between different population Σs and still the hypotheses related to the invariance of particular measurement or structural parameters might be rejected. Conversely, the equality of Σs may be rejected, yet tests for invariance of measurement and structural invariance may hold. Another test of invariance examines the Λ_x, Φ, and Θ_δ matrices. Typically this test is used as a baseline model against which the tenability of the other invariance models are tested.

To assess measurement equivalence, the regression coefficients or lambdas linking the items to their latent constructs are of primary interest (Dragow & Kanfer, 1985; Reise et al., 1993; Raju et al., 2000). The central issue in measurement equivalence is whether the Λs matrices are identical across populations (Dragow & Kanfer, 1985; Reise et al., 1993; Raju et al., 2000). A stricter definition of measurement equivalence may require item error variances also be equal across populations (i.e., $\sigma_{\delta_i}^2 = \sigma_{\delta_i}^2$ for all i). This would insure that the reliability of item scores are equivalent across populations, provided the variances of ξ s are equal across populations.

Some researchers (e.g., Little, 1997; Chan, 2000) argue that equivalence of the intercepts should be tested as well as the regression coefficients (lambdas) in examining measurement equivalence. This requires testing for the invariance of intercept structures as well as the lambda

structures, and is often referred to as the MACS analysis (Byrne, 1998). As previously noted, the intercepts in Equation 2 are assumed to be zero; an equation similar to Equation 2, with non-zero intercepts, may be expressed as:

$$x_i = \tau_i + \lambda_i \xi + \delta_i, \tag{7}$$

where τ_i is the intercept for item i. The expectation and variance of this equation may be expressed as:

$$\varepsilon(x_i) = \tau_i + \lambda_i \xi, \tag{8}$$

$$\sigma_{x_i}^2 = \lambda_i^2 \sigma_\xi^2 + \sigma_{\delta_i}^2. \tag{9}$$

The tau (intercept) may be expressed as:

$$\tau_i = \mu_\xi - \lambda_i \mu_{x_i} \tag{10}$$

It should be noted that, while their variances are identical, the expectations of Equations 2 and 7 differ by an intercept. Within this context (or within the MACS context), measurement equivalence may be defined as the equality of lambdas (regression coefficients or slopes) and taus (intercepts) across populations. As Raju et al. (in press) noted, the non-equality of taus may represent not only the differences in lambdas but also the differences in the means of x_i and ξ across populations. According to Equation 10, when the lambdas are equal, the difference in taus (intercepts) will simply reflect the difference in the means of x_i and ξ across populations. That is, the tau difference may be more a reflection of impact (Dorans & Holland, 1993) rather than DIF. Furthermore, any non-zero tau difference for an item across two populations does not vary as a function of ξ and hence represents a constant difference. This is one reason why some question the use of the tau difference as a measure of DIF (refer to Raju et al., in press). Despite

these concerns, the equality of lambdas only and the simultaneous equality of lambdas and taus are examined to get a better handle on the use of CFA for assessing measurement equivalence.

Differential Item Functioning (DIF) using Item Response Theory (IRT)

Many IRT-based measurement equivalence techniques, also called DIF procedures, are available for assessing measurement equivalence across populations. IRT-based measurement equivalence methods posit a nonlinear relationship between the latent construct and the observed score. Items are said to have measurement equivalence if the item parameters remain invariant across the two populations. The invariance of item parameters imply the invariance of item response functions (IRFs); that is, the assessment of measurement equivalence can be done either at the item parameter level or at the IRF level.

The following description of the IRT perspective uses the graded response model of Samejima (1969). Samejima's graded response model assumes an ordered response; that is, the more steps successfully completed, the larger the category score. The probability of person s responding above category k to item i is:

$$P_{ik}^{*}(\theta) = \frac{\exp[\,Da_i(\theta_s - b_{ik})]}{1 + \exp[\,Da_i(\theta_s - b_{ik})]}, \qquad (11)$$

where b_{ik} is the boundary or threshold between category k and $k + 1$ associated with item i; a_i is the item slope or discrimination parameter; and θ_s is the ability parameter. This equation is referred to as the boundary response function (BRF). The BRF is similar to the two-parameter dichotomous model except m-1 (m = number of categories in an item) functions are needed per item.

To calculate the probability of responding in a particular category, the adjacent boundary is subtracted from the cumulative probability. This can be expressed as:

$$P_{ik}(\theta) = P_{i(k-1)}^{*}(\theta) - P_{ik}^{*}(\theta). \qquad (12)$$

This function is often referred to as the item category response function (ICRF).

True Item Scores

Once the probability for responding in each category (i.e., ICRFs) is estimated, a measure of the item expected (raw) score could be calculated. For polytomously-scored data an expected score (ES_{si}) for item i can be computed for examinee s as:

$$ES_{si} = \sum_{k=1}^{m} P_{ik}(\theta_s) X_{ik}, \tag{13}$$

where X_{ik} is the score or weight for category k; m is the number of categories; and P_{ik} is the probability of responding to category k (see Equation 12). This is referred to as the item true score function (ITSF) or the item response function (IRF).

Definition of DIF or Measurement Equivalence

Chang and Mazzeo (1994) demonstrated that for the graded response model if two items have the same *ES* or *IRF* then they must have the same number of scoring categories and same item category response functions (*ICRF*). Conversely, an item is considered to be functioning differentially if:

$$ES_{iR} \neq ES_{iF}, \tag{14}$$

where ES_{iR} is the item expected score for an examinee in the reference group (R) (i.e., comparison group) with a given θ and ES_{iF} is the item expected score for an examinee in the focal group (F) (i.e., the group of interest) with the same θ for item i (see Equation 13).

IRT-based Procedure for Examining Measurement Equivalence or DIF

There are several IRT-based DIF procedures (e.g., Lord's chi-square (Lord, 1980), Raju's area measures (Raju, 1988, 1990), likelihood ratio test (Thissen, Steinberg, & Wainer,1988), and differential functioning of items and tests (DFIT) framework, and others). In this study we will

describe the DFIT framework (Raju, van der Linden, & Fleer, 1995) for assessing measurement equivalence.

The DFIT framework requires two separate item parameter estimations, one for the reference group and the other for the focal group. This results in two sets of item parameters for a test. The reference group distribution is used only for providing estimates of the item parameters for the group. The reference group item parameters are then placed on the same metric as the focal group parameters using a linear transformation. The focal group ability distribution is used to calculate two item expected scores (see Equation 13), one using the focal group parameters and the other using the transformed reference group parameters. That is, for a single examinee (with a given θ) who is a member of the focal group (F), an expected score for an item, ES_{siF}, can be calculated utilizing the focal group item parameters. For the same examinee (with the same θ), another expected score, ES_{siR}, is calculated using the reference group item parameters. If the item is functioning differentially, the two expected scores would not be equal (see Equation 14). Raju et al. refer to this DIF procedure as non-compensatory DIF (NC-DIF). The DFIT framework also has a procedure for detecting differential test functioning (DTF), but in this study only the DIF procedure (i.e., NC-DIF) will be examined. For a detailed description of the DFIT framework, refer to Raju et al. (1995).

Equivalence of the CFA and IRT Methods

Raju et al. (in press) examined the theoretical similarities and differences between the CFA and IRT perspectives in testing for invariance. Similarities of the methods are that both (a) examine the relationship between an underlying construct and a set of measured variables; (b) examine the degree to which item/subscale level true scores are similar for persons in the two populations with the same level of ability score on the latent construct; and (c) the definition of

measurement equivalence does not imply that the distributions of scores on the underlying constructs in the two populations of interest are identical. The differences between the two approaches are (a) the relations between the latent construct and the true score is linear in the CFA approach and non-linear in the IRT approach; (b) dichotomously scored measures are more appropriate for the IRT approach but as the number of possible scores for an item increase, the CFA approach will be equally tenable; and (c) CFA methodology is well advanced to handle multiple latent constructs and multiple populations simultaneously, whereas much of the IRT-based DIF is confined to unidimensional scales.

The purpose of this study was to evaluate the performance of CFA and IRT-based methods of examining measurement equivalence. Data were simulated and the measurement equivalence procedures were conducted on the simulated data.

Method

To evaluate the performance of the CFA procedures and an IRT-based procedure, a simulation study was conducted. This simulation represented a unidimensional 20-item test with all items having five-category responses. Item response data were generated using the graded response model (Samejima, 1969).

Data Simulation

A graded response model with 5-response categories was used to generate the simulated data sets. Item parameters used in previous studies (Flowers, Oshima, & Raju, 1997) were modified to accommodate the graded response model. The modified item parameters are contained in Tables 1 and 2. Note that DIF was modeled by adding (or subtracting) a constant to the *a*- and/or *b*-parameters of the focal group.

Next, probabilities for the five categories per item for a simulated examinee were generated using Equation 11. Recall that the five categories result in four probabilities per item. In order to assign a score for each simulated examinee, the following procedure was used. First, each simulated examinee was randomly assigned an ability parameter (θ) from a standard normal distribution. Using the item parameters in Tables 1 and 2 along with the randomly assigned ability parameter (θ), each simulated examinee has four probabilities per item. Next, for each simulated examinee, a single random number (X) was sampled from a uniform distribution over the interval [0,1]. If the randomly sampled number was less than the calculated probability at the boundary category k but greater than the calculated probability at $k + 1$, then the score assigned was the value of category k. This can be expressed as:

$$P^*_{sik} > X_{si} > P^*_{si(k+1)i}, \tag{15}$$

where X_{si} is the single random number for examinee s on item i.

Three conditions were simulated in this study by having different true item parameters for the focal and reference groups. In Condition 1, two of the 20 items were imbedded with DIF (Items 3 and 8) by adding a constant, .5 and 1.0, respectively, to the focal groups b-parameters. This type of DIF is typically referred to as uniform DIF. In Condition 2, four items (Items 3, 8, 13, and 18) were embedded with DIF; Item 3 had a constant added to the b-parameters (+1.0); Item 8 had a constant subtracted (-0.5) from the a-parameter and a constant added to the b-parameters (+0.5); for Item 13 a constant was added to the b-parameters (+0.5); and Item 18 had a constant subtracted (-0.5) from the a-parameter. In this condition both uniform DIF (i.e., differences in b-parameters) and non-uniform DIF (i.e., differences in a-parameters) was embedded. Condition 3 also had four DIF items (Items 3, 4, 12, and 13). In this condition, DIF was imbedded by adding and subtracting a constant from the a-parameters (Items 3 and 4) and

adding and subtracting from the *b*-parameters (Items 12 and 13). In the first two conditions, the reference group was favored; in the third condition, two items favored the reference groups and two items favored the focal groups.

Two ability distributions for the focal groups were simulated. For the first three simulated conditions (Conditions 1, 2, and 3), the focus groups had the same (normal) ability distribution as the reference groups (i.e., N(0, 1)). This is referred to as the no-impact scenario. In the second scenario, the focus groups' (normal) ability distributions were one standard deviation below the reference groups' distributions (i.e., N(-1, 1)). This is referred to as the impact scenario. The group sample sizes were 1000 examinees. Five replicated simulations were done for each of the three conditions and two scenarios.

Confirmatory Factor Analysis

For the CFA analyses, the software program LISREL (Jöreskog & Sörbom, 1996) was used to test for measurement equivalence across the focal and reference groups, first with lambdas (slopes) only and then with lambdas and taus (slopes and intercepts). Prior to the measurement equivalence analysis, in order to establish a baseline model, goodness-of-fit related to a one-factor CFA model was tested simultaneously across groups, with no equality constraints imposed. The χ^2 statistic related to this simultaneous model represents the sum of the χ^2 values resulting from test of model fit for each group separately. Next, the estimated values of the factor loading matrix (lambdas) were compared across the focal and reference groups. To test for the equality of lambdas, a comparison of fit between a model in which all factor loadings were constrained to be equal across groups and the baseline model was carried out. The difference in χ^2 values ($\Delta\chi^2$) for the two models is distributed as a χ^2, with degrees of freedom equal to the difference in degrees of freedom between the two models (Jöreskog & Sörbom, 1996). A

statistically significant (in this study $\alpha = .01$) χ^2 value would indicate noninvariance, and additional testing would be needed to identify which items in the scale were operating differently across the focal and reference groups. Models in which each item's lambda was constrained to be equal across groups were compared to the baseline model using the same statistical test ($\Delta\chi^2$). The statistically significant difference between models would indicate that the item was not operating equivalently across the focal and reference groups. The item level assessment of the equality of lambdas was sequential and inclusive in the sense that the previously identified items with measurement equivalence were constrained to have the same lambdas across the groups in the succeeding stages. Next, the same procedures were used for assessing the simultaneous equivalence of both lambdas and taus (slopes and intercepts) across the two groups.

Differential Item Functioning

Raju et al.'s (1995) procedure based on differential functioning of items and tests (DFIT) was used for the IRT-based DIF analyses. All item and ability parameters were estimated using the computer program PARSCALE 3 (Muraki & Bock, 1997). The estimation of equating coefficients was made by means of Baker's modified test characteristic curve method as implemented in the EQUATE 2.0 computer program (Baker, 1993). In this study, all parameter estimates for the reference group were equated to the underlying metric of the focal group. The χ^2 test for detection of DIF is very sensitive to sample size, so a cutoff criterion recommended by Raju (1999) was used. An item with 5 category response options is considered to have significant DIF if the χ^2 associated with the NCDIF index is statistically significant at the .01 level and the index itself is greater than .096.

Results

Results are reported in terms of true positives and false positives. Items that are true positive are DIF items identified as non-equivalent across the focal and references groups. False positive, or Type I errors, are non-DIF items that were identified as non-equivalent across the focal and references groups. Results are reported separately for the no-impact and impact scenarios. Within each scenario, information about true positives and false positives is presented for the CFA-based lambdas (only) comparison and lambdas and taus (combined) comparison and the IRT-based NC-DIF comparison, separately by condition.

No-Impact Scenario

True Positives. The true positive rates for the no-impact scenario are reported in Table 3. The numbers under the lambda, lambda/tau, and NC-DIF columns represent the proportion of times the items with imbedded DIF were correctly identified in five replications. For Condition 1, the lambda/tau and NC-DIF procedures successfully identified all items with imbedded DIF in all replications; however, the lambda procedure did not identify any of the DIF items accurately in any of the replications. Recall that these two items had DIF imbedded by adding a constant to the *b*-parameter.

In Condition 2, the lambda/tau and NC-DIF procedures identified Items 3, 8, and 13 correctly in all 5 replications. With respect to Item 18, which had imbedded DIF only in the *a*-parameter, the lambda/tau procedure identified the item accurately only once in 5 replications. The NC-DIF procedure failed to identify Item 18 accurately in all replications. The lambda procedure did better (in identifying items with imbedded DIF) in this condition than in Condition 1; however, the overall performance of this procedure was still not as good as that of the lambda/tau and NC-DIF procedures. For example, the hit rate for the lambda procedure was

perfect for Items 8 and 18 and was zero for Items 3 and 13, which had imbedded DIF only in the *b*-parameter.

In Condition 3, the lambda/tau and NC-DIF procedures identified Items 12 and 13 as having DIF in all replications. These items had DIF imbedded only in the *b*-parameters. The lambda/tau procedure accurately identified DIF in 3 out of 5 replications for Item 4 and in 2 out of 5 replications for Item 3. The NC-DIF procedure did not identify DIF for Items 3 and 4 in any of the five replications. On the other hand, the lambda procedure had a perfect detection rate for Items 3 and 4. With respect to Items 12 and 13, the lambda procedure had a zero hit rate across all five replications. It should be noted that, in this condition, Items 3 and 4 had imbedded DIF only in the *a*-parameters and Items 12 and 13 had imbedded DIF only in the *b*-parameters.

To better understand the CFA and NC-DIF measures of DIF, the lambda, tau, and NC-DIF indices for the focal and reference groups for one of the replications from Condition 3 are shown in Table 4. For Items 3 and 4, the lambdas are quite different between the focal and reference groups, while the tau coefficients appear to be very similar. For these two items, the NC-DIF indices are equal (.007) and less than the cutoff of .096. These are the items that the lambda procedure successfully identified as not having invariance in all replications, while and the lambda/tau and NC-DIF procedures were inconsistent in their identification. With respect to Items 12 and 13, the lambda procedure failed to detect non-invariance (or DIF), whereas the lambda/tau and NC-DIF procedures did. It appears that the lambda procedure is quite accurate when the differences are only in the a-parameters; similarly, the lambda/tau and NC-DIF procedures are very accurate when the differences are only in the b-parameters. We will have more to say about this finding later.

False Positives. For all the conditions in the no-impact scenario across all procedures, very few non-DIF items were falsely identified as exhibiting non-invariance (<.01 across all conditions and all procedures). For example, the false positive rates varied from .00 to .003 for the lambda procedure, from .00 to .004 for the lambda/tau procedure, and from .001 to .009 for the NC-DIF procedure across all conditions in the no-impact scenario. In other words, Type I error rates appear to fall in acceptable ranges for this scenario.

Impact Scenario

True Positives. The detection rates for DIF items for the different procedures for the impact scenario are reported in Table 5. Recall that the only difference between the no-impact and impact scenarios was that in the impact scenario the focal group (normal) distributions of thetas were one standard deviation below that of the reference group (normal) distributions of thetas. The true positive rates for this scenario were very similar to the true positive rates for the no-impact scenario. The lambda/tau and NC-DIF procedures detected non-invariance for all the items where there were differences in the *b*-parameters, but were inconsistent or not very successful in identifying items that had differences in only the *a*-parameters. It should be noted that the lambda/tau procedure had a slightly higher detection rate here for the nonuniform DIF items than in the no-impact scenario. The lambda procedure successfully identified all items with differences in the *a*-parameters but did not identify any items that had only differences in the *b*-parameters. The lambda procedure had identical results in the no-impact scenario (see Table 3).

False Positives. The major difference between the no-impact and impact scenarios was the difference in the Type I error rates. The lambda and NC-DIF procedures continued to have low Type I error rates in this scenario. For example, the false positive error rates for the impact scenario varied between .00 and .003 for the lambda procedure and between .003 and .01 for the

NC-DIF procedure, which were similar to the rates found in the no-impact scenario. However, the false positive error rates (proportions of times that non-DIF items were identified as having non-invariance) were substantially higher for the lambda/tau procedure. For example, the observed false positive error rates for the lambda/tau procedure under the impact scenario were .29, .27, and .43 for Conditions 1, 2, and 3, respectively.

Discussion

The findings of this study suggest that the type of DIF was a major factor for determining the detection rates of the different measurement equivalence procedures. When DIF was imbedded in the *b*-parameters, the lambda invariance procedure did not identify non-invariance (DIF) across the focal and reference groups; however, the lambda procedure was sensitive to differences in the *a*-parameters. Since the *a*-parameters in the IRT-based method are similar to the CFA regression coefficients (λ), we anticipated that the lambda procedure could detect nonuniform DIF.

The lambda/tau procedure successfully identified non-invariant items with differences in the *b*-parameters, but an unexpected finding was that it was not as sensitive to differences in the *a*-parameters only. Since MACS examines the means structures as well as the lambda structures of the focal and reference groups, it was anticipated that the lambda/tau procedure would identify uniform DIF as well as nonuniform DIF. Items with uniform DIF were identified as having non-invariance across focal and references groups; however, for the nonuniform items, the non-invariance detection rates decreased. When the difference between the focal and reference groups' ability distributions differed (i.e., impact), the lambda/tau procedure identified many non-DIF items as has having non-invariance. This observed Type I error rate was well above the α level.

The NC-DIF procedure successfully identified uniform DIF items, but was not as sensitive to nonuniform DIF items when there was difference only in the *a*-parameters. (The same phenomenon was also found for the lambda/tau procedure.) There may be an explanation for this occurrence. The NC-DIF index looks at the squared difference in the focal and reference group IRFs across the thetas of examinees in the focal group. The difference in the focal and reference group IRFs is simply the sum of the differences in the focal group and reference group BRFs. Each BRF difference is like a signed area, which, according to Raju (1988), depends only on the difference in the *b*-parameters. So, when there is a difference in only the *a*-parameters (and not in the *b*-parameters), the BRF difference will be close to zero, thus resulting in a very small or zero IRF difference. This will then lead to a non-significant NC-DIF index. Examples of small or zero true NC-DIF, when items differ only in the *a*-parameter, can be found in Flowers et al. (1999). It should be noted, however, that, when there are differences in both the *a*- and *b*-parameters, the NC-DIF procedure appears to be quite sensitive to DIF, as observed in the case of Item 8 in Condition 2.

To the best of our knowledge, this is the first Monte Carlo investigation of DIF (or measurement inequivalence) procedures based on CFA and IRT methodologies. While much is known about the IRT-based DIF procedures, our current knowledge about the CFA-based DIF procedures is quite limited. There is definitely a need for additional empirical and Monte Carlo investigations. According to the current results, however, the practical utility of the lambda procedure appears to be somewhat limited. The lambda/tau procedure appears to be promising in terms of true positives, but quite disappointing in terms of false positives. Considering both true and false positives, the IRT-based NC-DIF procedure may prove useful in practice.

When an item differs only in the *a*-parameter across focal and reference groups, the detection of DIF or measurement inequivalence appears to be a problem for all procedures. This may be more a function of how DIF is defined rather than the validity of the procedures involved. In both the CFA and DFIT frameworks, DIF is defined as the difference in true scores (or expected raw scores) for focal and reference group examinees with identical theta scores. As previously noted with respect to the NC-DIF procedure, when an item differs only in the *a*-parameter across the focal and reference groups, the item-level true scores may be quite similar and therefore may not represent DIF. Since most decisions are probably made at the item or test score level, an examination of differences in item parameters only may exaggerate DIF in a practical sense. This appears to be especially true for non-dichotomous items within the IRT-based framework. The question of when differences in item-level true scores and differences in item parameters converge in identifying DIF or measurement inequivalence needs to be investigated, especially with respect to non-dichotomous items.

The simulation in this study, which modeled the non-linear IRT measurement scale, favored the IRT-based method. Therefore, there is also a need for generating data with a CFA model and then assessing how well the CFA-based and IRT-based DIF procedures perform in assessing measurement equivalence. In addition, the number of conditions simulated was limited as well as the number of replications. Future studies should examine effects of samples sizes, number response categories per item, and differences in the magnitude of imbedded DIF to determine the sensitivity of the measurement equivalence procedures.

References

Baker, F. B. (1993). *EQUATE2: Computer program for equating two metrics in item response theory* [computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley.

Byrne, B.M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29*, 289-311.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Chan, D. (2000). Detection of differential item functioning on the Kirton adaption-innovation inventory using multi-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35(2)*, 169-199.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59*, 391-404.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. . In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*, 662-680.

Flowers, C.P., Oshima, T.C., & Raju, N.S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*, 309-326.

Jöreskog, K.J., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software.

Little, T.D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillside, NJ: Erlbaum.

Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks*. Chicago, IL: Scientific Software International.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207.

Raju, N.S. (1999). *DFIT5P: A Fortran program for calculating DIF/DTF* [Computer Program]. Chicago, Illinois Institute of Technology.

Raju, N.S., Laffitte, L.J., & Byrne, B.M. (in press). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*.

Raju, N.S., Laffitte, L.J., & Byrne, B.M. (April, 2000). *Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory*. Paper presented at the Annual Meeting of the Society for Industrial/Organizational Psychology, New Orleans.

Raju, N.S., van der Linden, W., & Fleer, P. (1995). An IRT-based internal measure of

test bias with applications for differential item functioning. *Applied Psychological Measurement, 19*, 353-368.

Reise, S.P., Widaman, K.F., & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement equivalence. *Psychological Bulletin, 114*, 552-566.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, No. 7.*

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Table 1

Reference Group Item Parameters Used to Simulate Data

Item	a	b_1	b_2	b_3	b_4
1	0.55	-1.80	-0.60	0.60	1.80
2	0.73	-2.32	-1.12	0.08	1.28
3	0.73	-1.80	-0.60	0.60	1.80
4a	0.73	-1.80	-0.60	0.60	1.80
4b	1.23	-1.80	-0.60	0.60	1.80
5a	0.73	-1.28	-0.08	1.12	2.32
5b	0.73	-1.80	-0.60	0.60	1.80
6a	1.00	-2.78	-1.58	-0.38	0.82
6b	0.73	-1.28	-0.08	1.12	2.32
7a	1.00	-2.32	-1.12	0.08	1.28
7b	1.00	-2.78	-1.58	-0.38	0.82
8	1.00	-2.32	-1.12	0.08	1.28
9a	1.00	-1.80	-0.60	0.60	1.80
9b	1.00	-2.32	-1.12	0.08	1.28
10a	1.00	-1.80	-0.60	0.60	1.80
10b	1.00	-1.80	-0.60	0.60	1.80
11	1.00	-1.80	-0.60	0.60	1.80
12a	1.00	-1.80	-0.60	0.60	1.80
12b	1.00	-1.28	-0.08	1.12	2.32
13a	1.00	-1.28	-0.08	1.12	2.32
13b	1.00	-0.78	0.42	1.62	2.82
14	1.00	-1.28	-0.08	1.12	2.32
15a	1.00	-0.82	0.38	1.58	2.78
15b	1.00	-0.82	0.38	1.58	2.78
16a	1.36	-2.32	-1.12	0.08	1.28
16b	1.00	-0.32	0.88	2.08	3.28
17a	1.36	-1.80	-0.60	0.60	1.80
17b	1.36	-2.32	-1.12	0.08	1.28
18	1.36	-1.80	-0.60	0.60	1.80
19	1.36	-1.28	-0.08	1.12	2.32
20	1.80	-1.80	-0.60	0.60	1.80

Note. [a]Item parameters used in Condition 1 and 2.
[b]Item parameters used in Condition 3.

Table 2

Focal Group Item Parameters Used to Simulate Data

Item	a	b_1	b_2	b_3	b_r	Difference in IP from Reference Group	
						a	b
	Condition 1						
3	0.73	-1.30	-0.10	1.10	2.30		+.5
8	1.00	-1.32	-0.12	1.08	2.28		+1.0
	Condition 2						
3	0.73	-0.80	0.40	1.60	2.80		+1.0
8	0.50	-1.82	-0.62	0.58	1.78	-0.5	+0.5
13	1.00	-0.78	0.42	1.62	2.82		+0.5
18	0.86	-1.80	-0.60	0.60	1.80	-0.5	
	Condition 3						
3	1.23	-1.80	-0.60	0.60	1.80	+0.5	
4	0.73	-1.80	-0.60	0.60	1.80	-0.5	
12	1.00	-0.78	0.42	1.62	2.82		+0.5
13	1.00	-1.28	-0.08	1.12	2.32		-0.5

Note. Focal group item parameters not include in this table are identical to the reference group item parameters in Table 1.

Table 3

Proportion of Items Identified as Displaying Inequivalence in the No-impact Scenario

	Differences in Item Parameters		CFA		
Items	a	bs	Λ	Λ/τ	NC-DIF
	Condition 1				
3	.0	+.5	.0	1.0	1.0
8	.0	+1.0	.0	1.0	1.0
	Condition 2				
3	0	+1.0	.0	1.0	1.0
8	-0.5	+0.5	1.0	1.0	1.0
13	.0	+0.5	.0	1.0	1.0
18	-0.5	.0	1.0	.2	.0
	Condition 3				
3	+0.5	.0	1.0	.4	.0
4	-0.5	.0	1.0	.6	.0
12	.0	+0.5	.0	1.0	1.0
13	.0	-0.5	.0	1.0	1.0

Note. The results are based on five replications.

Table 4

Estimated CFA Focal and Reference Groups' Lambda (λ), and Tau (τ); NC-DIF Value for One Replication in Condition 3(No-impact)

	CFA-Estimated Parameters				IRT-Index	Differences in IP	
	Focal	Reference	Focal	Reference			
Item	λ	λ	τ	τ	*NC-DIF*	a	b
1	0.451	0.508	2.00	1.94	.000		
2	0.550	0.558	2.32	2.32	.000		
3	**0.757**	**0.591**	**1.99**	**2.04**	**.007**	**0.5**	
4	**0.564**	**0.748**	**2.01**	**1.98**	**.007**	**-0.5**	
5	0.575	0.562	2.04	1.95	.000		
6	0.605	0.586	1.63	1.66	.000		
7	0.656	0.700	2.72	2.69	.000		
8	0.696	0.651	2.41	2.41	.000		
9	0.734	0.712	2.41	2.39	.000		
10	0.687	0.685	2.00	1.96	.000		
11	0.723	0.699	2.00	1.96	.000		
12	**0.708**	**0.686**	**1.17**	**1.56**	**.134**		**0.5**
13	**0.676**	**0.673**	**1.57**	**1.24**	**.134**		**-0.5**
14	0.688	0.705	1.61	1.61	.000		
15	0.699	0.671	1.32	1.30	.000		
16	0.661	0.686	.94	.96	.000		
17	0.761	0.787	2.38	2.42	.000		
18	0.797	0.808	2.02	1.98	.000		
19	0.792	0.803	1.58	1.59	.000		
20	0.871	0.871	1.98	1.99	.000		

Table 5

Proportion of Items Identified as Displaying Inequivalence in the Impact Scenario

	Differences in Item Parameters		CFA		
Items	a	bs	Λ	Λ/τ	NC-DIF
	Condition 1				
3	.0	+.5	.0	1.0	1.0
8	.0	+1.0	.0	1.0	1.0
	Condition 2				
3	0	+1.0	.0	1.0	1.0
8	-0.5	+0.5	1.0	1.0	1.0
13	.0	+0.5	.0	1.0	1.0
18	-0.5	.0	1.0	.4	.2
	Condition 3				
3	+0.5	.0	1.0	.4	.0
4	-0.5	.0	1.0	.8	.2
12	.0	+0.5	.0	1.0	1.0
13	.0	-0.5	.0	1.0	1.0

Note. The results are based on five replications.

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)